STOCK MARKET ANALYSIS

A MINI-PROJECT REPORT

submitted by

JOEL T MATHEW (SCM21CS070)
MAGHNUS RICHARD (SCM21CS076)
MATHEW C PAUL (SCM21CS079)
NANDANA P (SCM21CS092)

to

The APJ Abdul Kalam Technological University

in partial completion of the criteria for the degree award

of

Bachelor of Technology

Department of Computer Science and Engineering

SCMS SCHOOL OF ENGINEERING AND TECHNOLOGY

(*Affiliated to APJ Abdul Kalam Technological University*)

VIDYA NAGAR, PALISSERY, KARUKUTTY

ERNAKULAM – 683576

MAY 2024

DECLARATION

We undersigned hereby declare that the project report Stock Market Analysis, submitted for partial fulfillment of the requirements for the award of degree of Master of Technology of the APJ Abdul Kalam Technological University, Kerala is a bonafide work done by us under supervision of Dr. Manish TI. This submission represents our ideas in our own words and where ideas or words of others have been included, we have adequately and accurately cited and referenced the original sources. We also declare that we have adhered to ethics of academic honesty and integrity and have not misrepresented or fabricated any data or idea or fact or source in my submission. We understand that any violation of the above will be a cause for disciplinary action by the institute and/or the University and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been obtained. This report has not been previously formed the basis for the award of any degree, diploma or similar title of any other University

Karukutty
09-05-2024

Joel T Mathew
Maghnus Richard
Mathew C Paul
Nandana P

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SCMS SCHOOL OF ENGINEERING AND TECHNOLOGY

(*Affiliated to APJ Abdul Kalam Technological University*)

VIDYA NAGAR, PALISSERY, KARUKUTTY

ERNAKULAM - 683582

CERTIFICATE

This is to certify that the report entitled 'Stock Market Analysis' submitted by **Joel T Mathew, Maghnus Richard, Mathew C Paul, Nandana P** to the APJ Abdul Kalam Technological University in partial fulfillment of the requirements for the award of the Bachelor of Technology is a bonafide record of the project work carried out by her under my guidance and supervision.

Project Guide	Project Coordinator	Head of the Department
Dr. Manish T I Professor, Dept. of CSE, SSET	Josna Philomina Asst. Professor, Dept. of CSE, SSET	Dr.Manish TI Professor, Dept. of CSE, SSET

ACKNOWLEDGEMENT

We would like to express our gratitude towards Dr. Anitha G Pillai, Principal, SCMS School of Engineering and Technology, Ernakulam and Dr. Manish TI, Head of Department, Department of Computer Science and Engineering, SCMS School of Engineering and Technology, for providing the resources and opportunities that made this project possible. This project has been a valuable learning experience, and we are grateful for the knowledge and skills gained during our time at this institution.

We are extremely grateful to our guide, Dr. Manish TI and our Project coordinator, Ms. Josna Philomina for their invaluable guidance, mentorship, and continuous support throughout the project. Their expertise and insightful feedback greatly enhanced the quality of our work. We extend our thanks to the faculty members of the Department of Computer Science and Engineering, SCMS School of Engineering & Technology, for their efforts in imparting knowledge and skills that have been instrumental in shaping our project. Their commitment to excellence and passion for teaching has been truly inspiring.: We would like to extend our appreciation to our friends and family for their valuable discussions, feedback, and collaboration during the course of this project. Their insights and constructive criticism have been instrumental in shaping our ideas and refining our work. Thank you to everyone who played a part in this project and contributed to its success. Your support has been invaluable, and we are truly grateful for your involvement

Joel T Mathew

Maghnus Richard

Mathew C Paul

Nandana P

ABSTRACT

The Stock Market Predictor with Sentiment Analysis mini-project addresses a pressing need within the financial realm, aiming to provide a sophisticated solution for accurate stock price forecasting by integrating sentiment analysis. Targeting the opening prices of five major companies—TCS, Infosys, Wipro, Reliance, and HDFC Bank—the project harnesses a dual data approach, sourcing historical market data from Yahoo Finance Library and conducting sentiment analysis through web scraping of financial news articles from Moneycontrol. Employing LSTM (Long Short-Term Memory) machine learning models, the system not only predicts stock prices based on historical trends but also factors in market sentiment gleaned from news sentiment analysis. The entire system is seamlessly hosted on a local server, ensuring real-time analysis and accessibility. Developed using Python, with Streamlit serving as the framework for deploying ML models, this mini-project represents a fusion of cutting-edge technologies and methodologies. Its overarching aim is to empower traders and investors with comprehensive insights, blending predictive analytics with nuanced market sentiment interpretations, all readily available through a user-friendly, locally hosted web interface.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ML	Machine Learning
LSTM	long short-term memory
JSON	JavaScript Object Notation
IDE	Integrated Development Environment
RMSE	Root Mean Squared Error
Adam	Adaptive Moment Estimation
Bert	Bidirectional Encoder Representations from Transformers(sentiment analysis)
tanh	hyperbolic tangent(activation function)

CHAPTER-1

INTRODUCTION

In the rapidly evolving world of finance, the ability to accurately predict the movement of stock prices has become a highly sought-after skill. Investors, traders, and financial institutions alike are constantly seeking ways to gain an edge in the market, and one of the most promising approaches is through the use of machine learning techniques. These techniques offer the potential to analyze vast amounts of data and identify complex patterns that may influence stock price movements, thus providing valuable insights for decision-making.

This project aims to explore the potential of these techniques in predicting stock market prices and understanding market sentiment by leveraging historical stock price data and sentiment analysis of news articles. By combining quantitative analysis of historical stock data with qualitative insights from sentiment analysis, the project seeks to develop a comprehensive understanding of the factors driving stock price movements. This holistic approach not only enhances the accuracy of stock price predictions but also provides valuable context for interpreting market dynamics and making informed investment decisions.

We are developing a predictive Machine Learning model using LSTM that can provide valuable insights into the future stock opening price of five companies: TCS, Infosys, Wipro, Reliance, and HDFC Bank. By training the model on historical stock data and integrating sentiment analysis of news articles, we aim to create a powerful tool for traders and investors to anticipate market trends and make data-driven decisions. Through this project, we hope to contribute to the advancement of predictive analytics in finance and empower stakeholders with actionable insights for navigating the complexities of the stock market.

CHAPTER-2

Literature Review

2.1 Inference

Deep Learning Enhances Prediction Accuracy:

Traditional machine learning techniques like Linear Regression and Support Vector Regression (SVR) have limitations in accurately predicting stock prices due to the non-linear and volatile nature of the stock market. Deep learning techniques, particularly Long Short-Term Memory (LSTM) networks, offer improved accuracy in predicting stock price movements by capturing time-series dependencies effectively.

A neural network set up using LSTM is compared with Support Vector Regression. Experiment is conducted on different stock indexes such as S&P 500, NYSE, NSE, BSE, Dow Jones Industrial Average, and NASDAQ. Experiment analysis proves that LSTM outperforms SVR and provides better prediction accuracy.

Relevance of Model Architecture:

The deployment of LSTM with specific configurations, such as multiple hidden layers, Adam optimizer, sigmoid activation function, and dropout, contributes to its superior performance in stock price prediction. Additionally, the use of appropriate libraries like Keras, Scikit-learn, Pandas, Numpy, and Matplotlib ensures efficient implementation and experimentation

Generalizability Across Multiple Indices:

The effectiveness of LSTM is demonstrated across various stock indexes such as NSE, BSE, NASDAQ, NYSE, S&P 500, Dow Jones Industrial Average, and Nikkei 225. This indicates the robustness and generalizability of LSTM in predicting stock price movements across different markets.

The literature review has led us to choose LSTM and Random Forest Regression as two machine learning models to compare against and choose the best in terms of accuracy for the final presentation.[2]

The inference drawn from the passage regarding random forest regression is that it is a promising technique for various financial applications, including stock market forecasting. Studies have shown that random forest regression performs well in predicting financial outcomes and market movements compared to other traditional methods like logistic regression or ordinary linear regression.

Specifically, Creamer and Freund utilized random forest regression to predict performance and quantify corporate governance risk in Latin American markets, showing favorable results compared to logistic regression.

While experimenting with random forest regression we found it to have very high accuracy but we found a fundamental flaw in its algorithm - if the stock price goes above the trained dataset values it can’t extrapolate the data ie; can't predict values outside its trained dataset. Therefore we chose LSTM as our Machine Learning model.[1]

CHAPTER-3

Problem Statement

The unpredictability of stock prices has always been a significant issue for investors and financial analysts. Traditional methods of stock price prediction, such as fundamental analysis and technical analysis, have limitations due to the influence of unpredictable factors like political events and market sentiment. Therefore, there is a need for a more sophisticated model that can capture the non-linear dependencies and patterns in stock price data. There are multiple instances where people lose money and even risk the financial well- being of the entire family. These instances arrive due to the unfamiliarity and lack of sound knowledge of the market. To combat this we have used Machine Learning and Sentiment Analysis to predict the values of the stocks thereby giving people confidence in the stocks they have invested.

Objective

Our project's primary objectives revolve around elevating the predictive capabilities in financial markets through Machine Learning

Predictive Modeling: To develop predictive models for forecasting future stock prices based on historical data using LSTM

Sentiment Analysis: We are also using news data of companies for doing sentiment analysis on news headlines and incorporating it with LSTM

User Interface: To create an interactive user interface using Streamlit, allowing users to select a company and view its stock market prediction.

Our project focuses on 5 specific stocks - TCS, Infosys, Wipro, Reliance, and HDFC Bank,

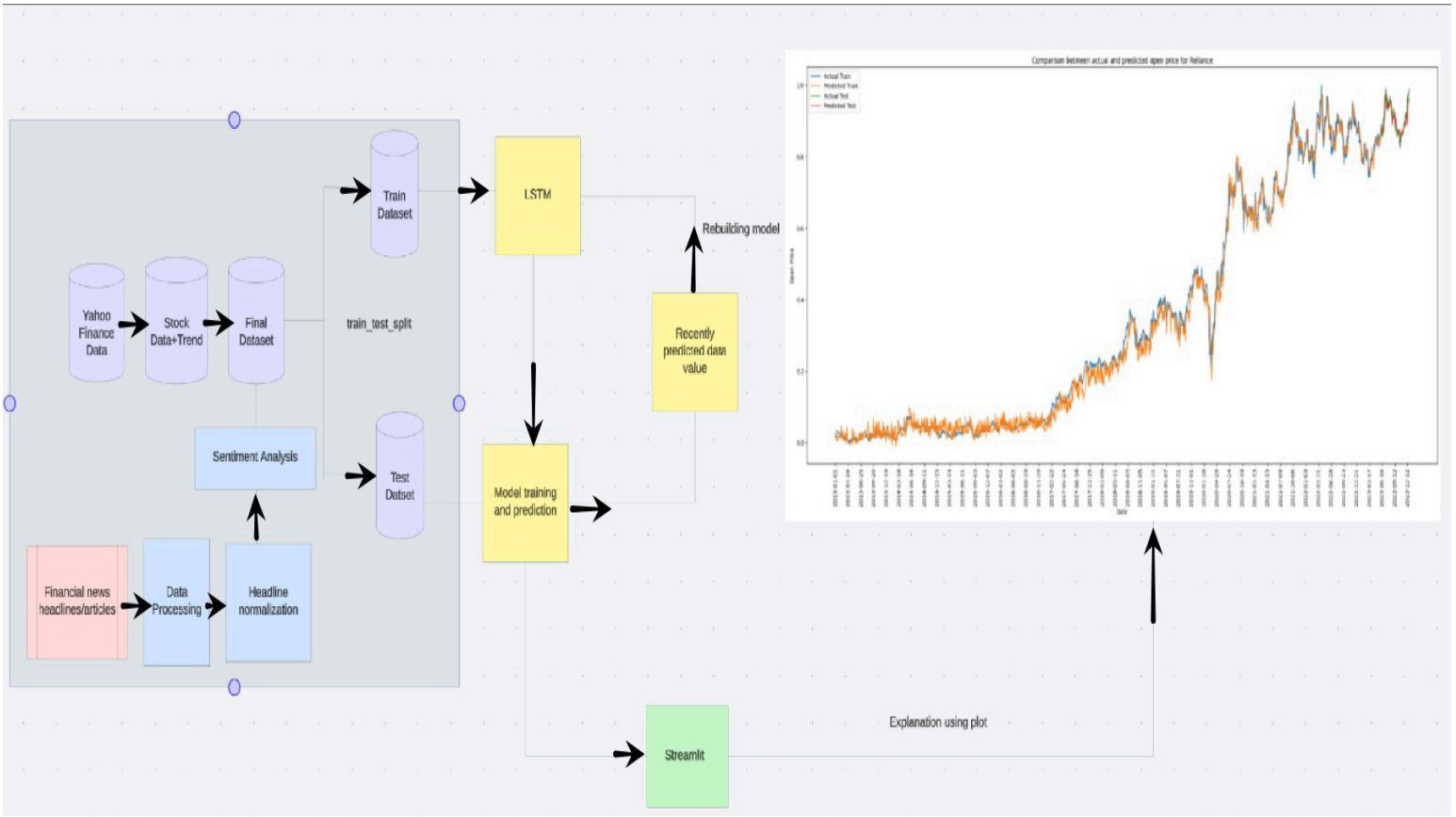
FUTURE SCOPE

- Including more companies(eg: like all companies in an index like nifty50)

CHAPTER-4

System Description

System Architecture



This architecture shows how the Machine learning Model is trained by using the datasets derived.

Fig 4.1 System Block Diagram

4.1.1 News Sentiments

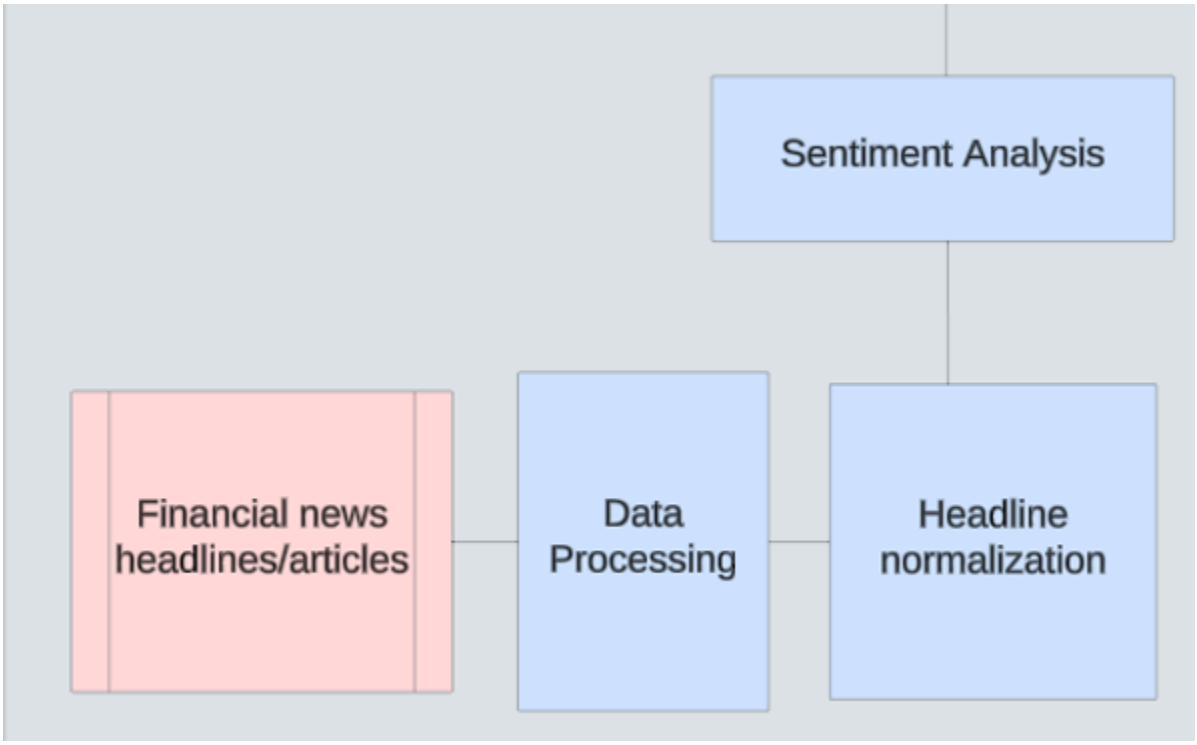


Fig 4.2 News Sentiment

The news headlines are obtained by web scraping data from a moneycontrol website and the headlines are then put through a Sentiment analyser to determine the sentiment of the news headline.

4.1.2 Stock Price

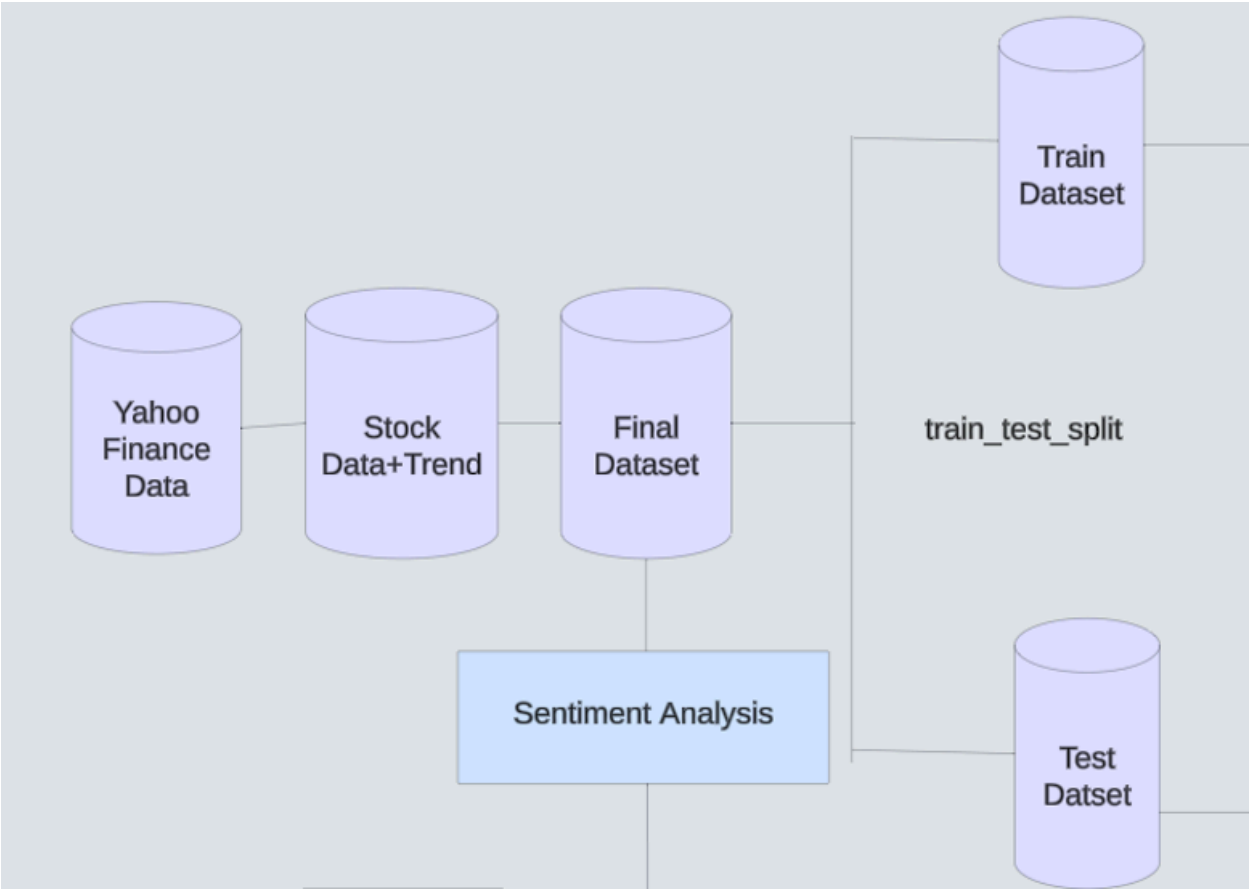


Fig 4.3 Stock Price

The daily open stock prices of each stock is obtained by using the YFinance module and paired with the sentiment data obtained above.

The entire dataset is then split 95:5 as Train and Test datasets respectively.

4.1.3 Model creation

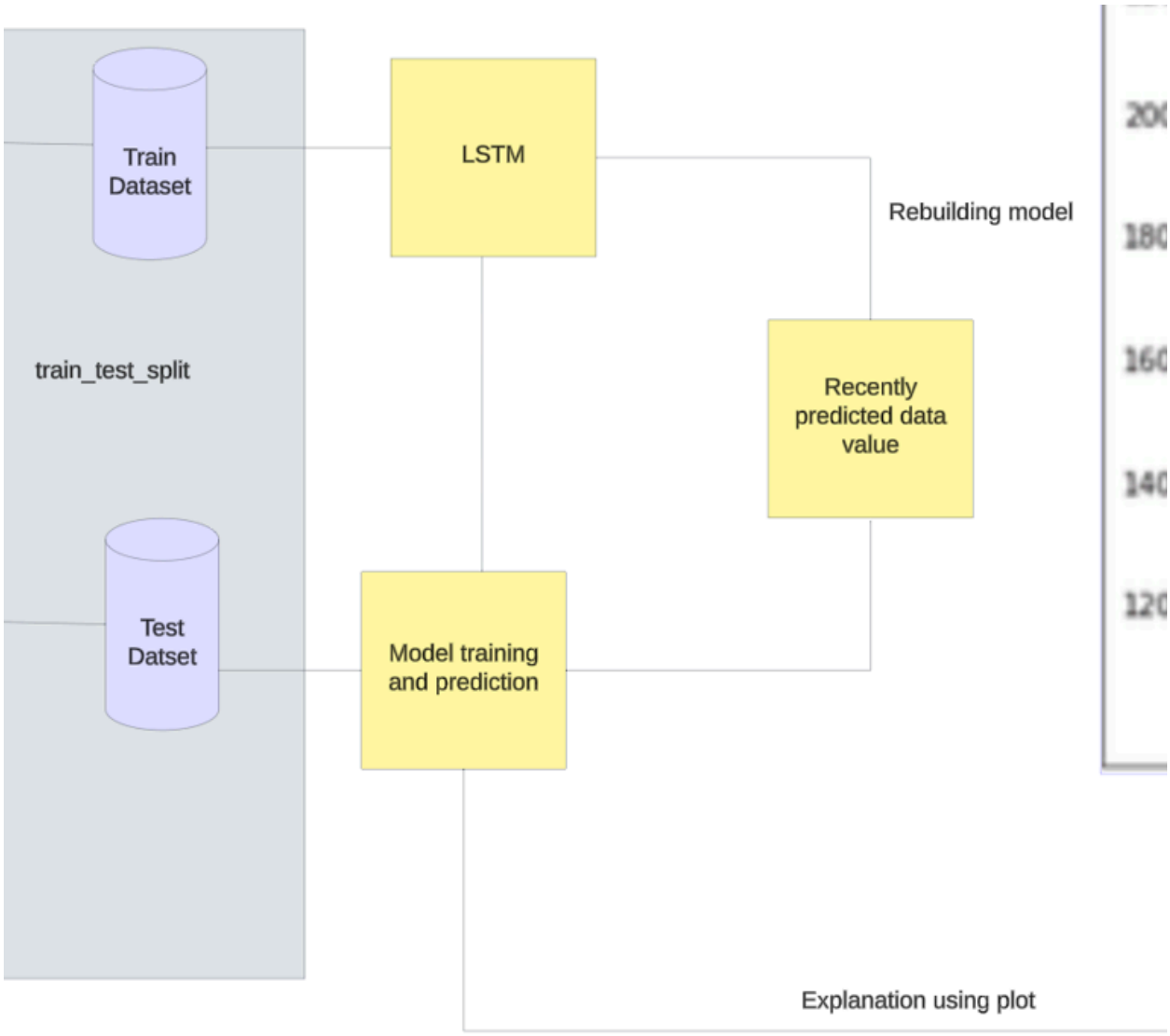


Fig 4.4 Model Creation

Based on the train dataset we train the LSTM model and we use the test dataset to predict and test accuracy of the model, we represent the results through some plots in the graph.

CHAPTER-5

System Design

Software

5.1.1Tools Used

The Application is based on:

ML Model: LSTM

Sentiment model: nlptown/bert-base-multilingual-uncased-sentiment

Web Scraping: Beautifulsoup 4, Requests

Text Editor: VSCode

Languages: Python

Operating System: The deep learning frameworks are compatible with different operating systems, including Windows, Linux, and macOS

Hosting: Streamlit

5.1.2 Algorithm

Step 1: Import the yfinance module in Python.

Step 2: Use the yfinance module to retrieve 10 years of historical data for the specified stocks and save it into a CSV file named stock.csv.

Step 3: Scrape 10 year news data from the Money Control website for the specified five companies.

Step 4: Save the scraped data into a CSV file named news.csv.

Step 5: Perform sentiment analysis on the news.csv file to analyze the sentiment of the news articles.

Step 6: Merge the output of the sentiment analysis onto the stock.csv file, and put it in a new csv file called merged.csv

Step 7: Preprocess the merged dataset for training an LSTM model.

Step 8: Scale the dataset to values between 0 and 1.

Step 9: Split the preprocessed dataset into training and testing sets with a ratio of 95:5.

Step 10: Create trainX, testX, trainY, and testY arrays for the LSTM model.

Step 11: Build the LSTM model using the reprocessed dataset as input.

Step 12: Train the LSTM model on the training data.

Step 13: Save the trained LSTM model to a .h5 file for later use in the Streamlit(web framework).

Step 14: In Streamlit, load the saved LSTM model.

Step 15: Use the loaded model and the same variables used in building the LSTM model to plot the graph of actual versus predicted open prices.

Step 16: Perform evaluation metrics like RMSE and display the accuracy of the model.

Step 17: Display a table that shows the actual open prices and the predicted open prices.

5.1.3 Dataset

Utilized stock data from TCS, Infosys, Wipro, Reliance, and HDFC Bank using YFinance python module.

Integrated news headlines through web scraping.

We used Moneycontrol to web scrape news data of the stocks.

Paired sentiment analysis data with daily open stock prices.

screenshots are data for a single company- HDFC bank, like that we have for all 4 remaining companies mentioned above

STOCK DATA

Date	Open	High	Low	Close	Volume
2013-01-01	341.05	342.55	339.83	342.25	2014606
2013-01-02	344.95	345	341.52	343.67	4854798
2013-01-03	345	345	340.2	341.67	4571366
2013-01-04	342.5	342.5	336.4	339.67	5442254
2013-01-07	341.85	341.85	333	334.1	5502284

Table 5.1

NEWS DATA

Scraped 10-year news data from moneycontrol website

Date	News
23 Aug 2019	business/hdfc-bank-to-offer-2-mn-creditdebit-cards-to-millennialsnext-2-years
12 Sep 2019	business/hdfc-bank-doubles-mid-corporate-loan-book3-years-may-cross-rs-1l-cr-mark-by-sept-end
13 Sep 2019	stocks-views/podcast-|-stock-picksthe-day-nifty-likely-to-tradethe-range10740–11100
23 Sep 2019	buy-hdfc-bank-targetrs-1510-sharekhan
26 Sep 2019	business/hdfc-bank-emerges-as-india39s-most-valuable-brand

Table 5.2

AFTER SENTIMENT ANALYSIS

The output you're seeing is the result of sentiment analysis using the `nlptown/bert-base-multilingual-uncased-sentiment` model.

- The model returns a list of probabilities for each of the five sentiment classes: 1 star, 2 stars, 3 stars, 4 stars, and 5 stars. These classes represent the sentiment of the text from very negative (1 star) to very positive (5 stars).

In your sample output:

- "04 Jan 2011" and "07 Jan 2011" are the dates for which the sentiment analysis was performed.

- The list of five numbers under each date represents the probabilities of each sentiment class for the text analyzed on that date.

For example, on "04 Jan 2011", the model predicted:

- 14.24% probability that the sentiment is very negative (1 star)

- 19.98% probability that the sentiment is negative (2 stars)

- 31.16% probability that the sentiment is neutral (3 stars)

- 21.83% probability that the sentiment is positive (4 stars)

- 12.79% probability that the sentiment is very positive (5 stars)

These probabilities tell you how likely the model thinks each sentiment is for the text it analyzed. The sentiment with the highest probability is the model's final prediction. In this case, the model predicts a neutral sentiment for "04 Jan 2011" as it has the highest probability (31.16%).

MERGING STOCK DATA AND SENTIMENT DATA(the feature1-5 shows the result of sentiment analysis)

merged data

Date	Open	High	Low	Close	Volume	Feature1	Feature2	Feature3	Feature4	Feature5
2019-07-22	1172.5	1174.93	1140.63	1148.63	11085648	0.15	0.12	0.24	0.25	0.25
2019-07-23	1147.25	1147.25	1121.03	1131.75	13901432	0.14	0.09	0.16	0.24	0.37
2019-07-24	1128.38	1146.5	1125.5	1140.45	8681870	0	0	0	0	0
2019-07-25	1142.5	1157.72	1140.13	1143.03	9731004	0	0	0	0	0
2019-07-26	1139.1	1146.82	1136.43	1138.65	6414914	0	0	0	0	0
2019-07-29	1136.78	1136.78	1111.53	1122.15	4988246	0	0	0	0	0
2019-07-30	1127.5	1135.6	1119.95	1126.13	5197068	0.16	0.18	0.24	0.24	0.18

Table 5.3

Conduct of Experiment/Working of the project

Here's how the project would work based on the outlined steps:

1. Data Collection

- Historical stock data for the past 10 years is retrieved using the `yfinance` module. This data includes attributes such as open, high, low, close prices, and volume.

- 10 years news articles related to the specified companies are scraped from the Money Control website.

2. Data Preparation:

- The scraped news data is saved into a CSV file for further processing.

- Sentiment analysis is performed on the news articles to determine the sentiment (positive, negative, or neutral) associated with each article.

- The sentiment analysis output is merged with the historical stock data.

3. Data Preprocessing:

- The merged dataset is preprocessed to prepare it for training the LSTM model.

- Data scaling is performed to normalize the values between 0 and 1, which is necessary for neural network training.

- The preprocessed dataset is split into training and testing sets, with 95% of the data used for training and 5% for testing.

4. Model Building:

- An LSTM (Long Short-Term Memory) model is built using the Keras library, a popular deep-learning framework in Python.

- The LSTM model is designed to learn patterns and relationships in the historical stock data and sentiment scores.

-The LSTM network consists of four layers with 100, 80, 50, and 30 units respectively. Dropout regularization is applied after each layer to prevent overfitting.

5. Model Training:

- The LSTM model is trained using the training data prepared earlier.
- During training, the model adjusts its internal parameters to minimize the difference between the actual and predicted stock prices.

6. Model Evaluation:

- Once the model is trained, it is evaluated using the testing data.
- Evaluation metrics such as Root Mean Squared Error (RMSE) are calculated to assess the model's performance.
- The accuracy of the model is also calculated to determine how well it predicts the stock prices.

7. Visualization and Reporting:

- The trained LSTM model is saved to a `.h5` file for later use.
- In a Streamlit web application, the saved model is loaded.
- Using the loaded model, actual and predicted stock prices are plotted on a graph to visualize their comparison.
- Evaluation metrics and accuracy are displayed to provide insights into the model's performance.
- A table showing actual and predicted stock prices is presented to the user for further analysis.

We did extensive testing on hyperparameter tuning for LSTM model.

This workflow demonstrates the end-to-end process of building a predictive model for stock price forecasting using LSTM and sentiment analysis, along with visualization and reporting using Streamlit.

CHAPTER-6

Results

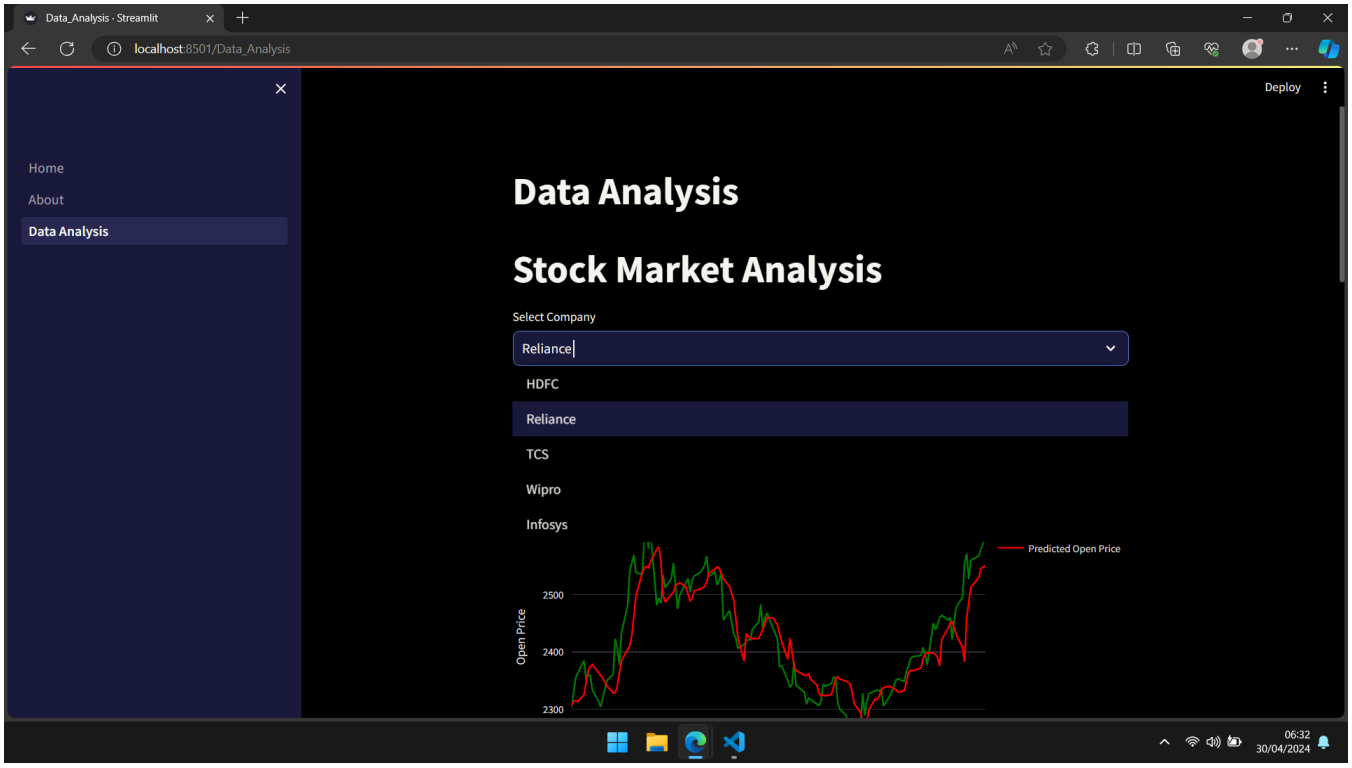


Figure 6.1 Data analysis page, which shows a dropdown menu of all companies

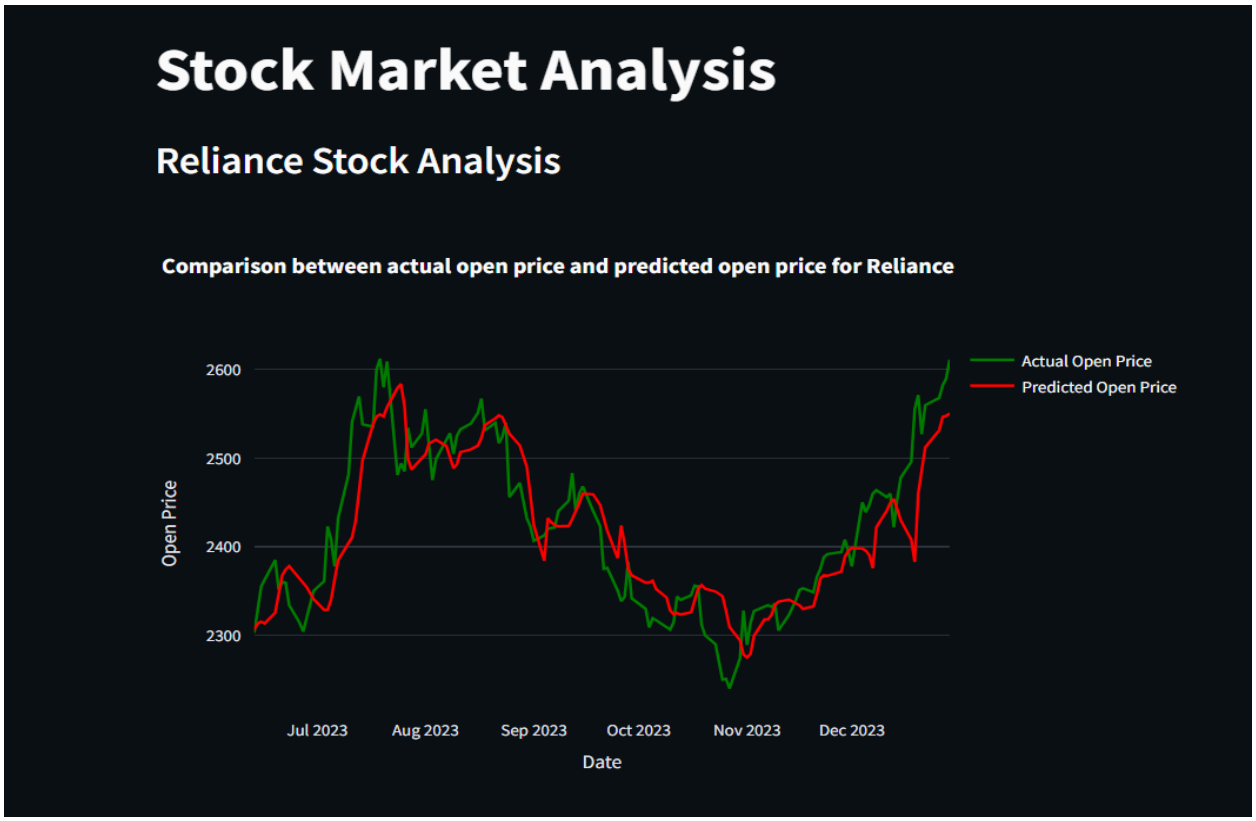


Figure 6.2 graph predicting opening price of reliance

Actual vs Predicted Open Price for Reliance:

	dates	Actual_Open_price	Predicted_Open_price
0	2023-06-13	2,302.8843	2,306.437
1	2023-06-14	2,326.8821	2,313.3533
2	2023-06-15	2,355.3105	2,315.2368
3	2023-06-16	2,363.0637	2,313.4497
4	2023-06-19	2,385.0313	2,325.6064
5	2023-06-20	2,351.3879	2,350.4224
6	2023-06-21	2,360.1101	2,367.709
7	2023-06-22	2,359.1873	2,374.4143
8	2023-06-23	2,334.1738	2,378.1123
9	2023-06-26	2,313.9602	2,363.9548

Figure 6.3 Predicted and Actual Opening price value table of reliance

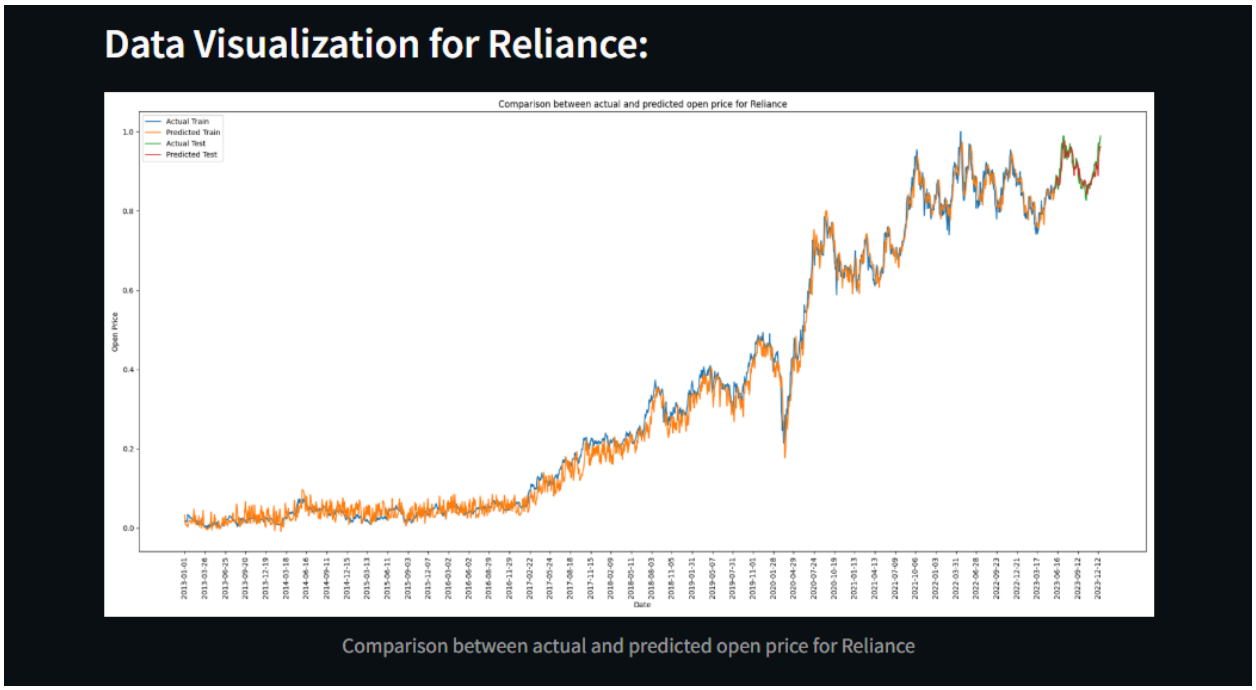


Figure 6.4 Graph showing train,test and actual opening price of reliance

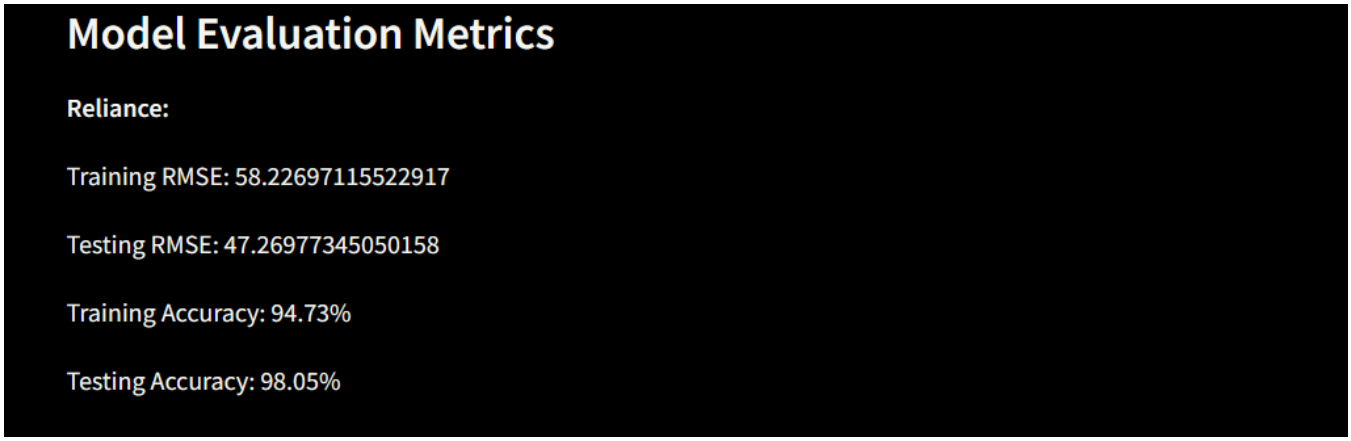


Figure 6.5 Model evaluation metrics

The above are the screenshots for reliance.

The remaining companies can be found by using the **drop down menu shown in Figure 6.**1

CHAPTER-7
Conclusion & Future Scope

Conclusion

Our project provided valuable insights into financial market dynamics, trends, and opportunities.
The integration of LSTM and sentiment analysis significantly improved predictive capabilities.
Investors and professionals can confidently navigate the market with enhanced accuracy and risk management strategies.

Future Scope

Implementation of all stocks in the market.
Advanced Techniques: Explore advanced prediction methods like LSTM-SVR integration.
Market Expansion: Scale analysis to include more stocks and markets.
Data Integration: Incorporate additional data sources like social media sentiment for richer analysis.
Tool Development: Create user-friendly tools for real-time decision support.
Research Focus: Continue research for publication, exploring emerging trends in stock market analysis.

CHAPTER-8

References

1. Stock Price prediction using LSTM and SVR-Published in: 2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC)-barthla G
   Date Added to IEEE Xplore: 04 December 2017
   DOI: 10.1109/PDGC50313.2020.9315800
   Publisher: IEEE
   https://sci-hub.se/10.1109/PDGC50313.2020.9315800
2. Forecasting Stock Index Movement: A Comparison of Support Vector Machines and Random Forest-Manish Kumar,M. Thenmozhi, Indian Institute of Technology Madras
   Indian Institute of Capital Markets 9th Capital Markets Conference Paper
   published in-https://papers.ssrn.com/sol3/papers.cfm?abstract_id=876544
   https://dx.doi.org/10.2139/ssrn.876544

3. Stock price prediction using LSTM on Indian share market-
   Stock-Price-Prediction-Using-LSTM-on-Indian-Share-Market.pdf (researchgate.net)
4. analyticsvidhya- stock price prediction using LSTM
   https://www.analyticsvidhya.com/blog/2021/12/stock-price-prediction-using-lstm/
5. towards Data Science- Machine Learning to Predict Stock Prices | by Roshan Adusumilli | Towards Data Science
6. Python libraries for beautifulsoup4-beautifulsoup4 · PyPI
7. Scikit-learn Documentation. (2022). Scikit-learn Documentation

   https://scikitlearn.org/stable/documentation.html Psycopg Documentation. (2022).

8. Plotly Python Graphing Library. https://plotly.com/python/
9. LSTM-LSTM layer (keras.io)
10. Streamlit Documentation https://docs.streamlit.io/